DOCUMENT RESUME

ED 274 700 TM 860 577

AUTHOR McLarty, Joyce R.
TITLE Tennessee's Career Ladder: How Far to Space the Rungs.
PUB DATE Apr 86
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Career Ladders; Competence; Elementary Secondary Education; Evaluation Criteria; *Evaluation Methods; *Job Performance; Merit Pay; *State Programs; *Teacher Evaluation
IDENTIFIERS Tennessee

ABSTRACT

In Tennessee, varied data-gathering devices have been employed in an attempt to determine whether an educator is qualified for a given level on the state's career ladder program and, thereby, should attain the meaningful financial benefits linked to that status. In its first year of operation, the Tennessee program employed several domains of competence, e.g., classroom management, and established per-domain performance standards which could be attained as consequence of examinees' performances on varied measures. The stages in the development of these standards are described, as well as the rationale underlying the standard-setting process. Information on results of the standards and changes made in them after the first year are included. Guidelines for those entrusted with this responsibility are provided. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED274700

TENNESSEE'S CAREER LADDER:

HOW FAR TO SPACE THE RUNGS

Joyce R. McLarty

Department of Education
State of Tennessee

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. R. McLarty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented to:

American Educational Research Association
and
National Council on Measurement in Education

San Francisco, California
April, 1986

Running Head: TN CAREER LADDER

TM 860 577

ABSTRACT

Tennessee's Career Ladder:

How Far to Space the Rungs

In Tennessee, varied data-gathering devices have been employed in an attempt to determine whether an educator is qualified for a given level on the state's career ladder program and, thereby, should attain the meaningful financial benefits linked to that status. In its first year of operation, the Tennessee program employed several domains of competence, e.g., classroom management, and established per-domain performance standards which could be attained as consequence of examinees' performances on varied measures.

The stages in the development of these standards are described, as well as the rationale underlying the standard-setting process. Information on results of the standards and changes made in them after the first year are included. Guidelines for those entrusted with this responsibility are provided.

Tennessee's Career Ladder:

How Far To Space The Rungs

Tennessee's Career Ladder program is a statewide system designed to pay teachers more for teaching well. Because it must make summative determinations, the Career Ladder program is a high-stakes enterprise. For that reason, many efforts were made to reduce the risk level for teachers entering the new Career Ladder program. The entire program was made optional for teachers currently employed when it was instituted. Those who did not wish to enter the Career Ladder were permitted to keep and to renew their current teaching certification under the former rules and regulations. For those willing to enter the new Career Ladder program, the "toe in the water" provision allowed a one time only return to the former certification system at any time.

During the first year of implementation, 1984-85, entry to Level I, the third of the five levels of the Career Ladder (the first and second are the probationary and apprentice levels), was made as risk free as possible. Teachers and administrators were offered five methods of entering Level I: exceed a minimum score on the National Teacher Exam Core Battery, exceed a minimum score on one of the National Teacher Exam Subject Area Tests, exceed a minimum score on the state-developed Career Ladder Test, satisfactorily complete an approved staff development program, or successfully undergo an approved local evaluation. Advancement to Level I of the Career Ladder carried with it a $1,000 annual supplement, plus retirement contributions for five years. More than 92% of the eligible educators in the state qualified for Level I.

Efforts were also made to reduce the risk level for teachers wishing to try for the upper two levels of the Career Ladder, Level II and Level III. A comprehensive orientation manual was provided. Teachers were encouraged to take the required written test as often as it was offered (up to five times) and permitted to use their highest score. Candidates were allowed to withdraw from or delay the evaluation process at any time without penalty. Candidate evaluation records were kept confidential throughout the evaluation process. Candidates were encouraged to see the evaluation process as the first step in a life-long career development program where everyone could continue to improve and no one was to be considered a failure.

Evaluation for the upper levels of the Career Ladder was a situation in which candidates had a great deal to gain. For

those who qualified, it meant additional money (up to $7,000 per year for up to five years). The prestige of being identified as a meritorious teacher was also an important benefit.

Success and failure mean different things to different people. To the individual teacher-candidate, success means qualifying for the highest level of the Career Ladder. For most of the rest of the state it meant finding and rewarding the best teachers in the state in order to attract and retain them as Tennessee teachers. The latter goal was well understood by those who developed and monitored the standards for the Tennessee Career Ladder program, and it guided the entire process.

The nature of the standard-setting endeavor was dictated in part by the evaluation system itself and by the guidelines established for it. The Career Ladder evaluation system is a competency-based system in which information is collected and then aggregated from multiple data sources. There are six domains of competence: preparation for instruction, teaching strategies, evaluation, classroom management, leadership, and communications. Communications is a screening (pass/fail) competency; the other competencies are scored. In 1984-85, each competency was measured by combining information from three or more of the data sources: classroom observation, portfolio analysis, principal questionnaire, student questionnaire, peer questionnaire, written test, and evaluator judgment.

Standard-setting was viewed as an integral part of the Career Ladder development process, rather than as something extra to be "added on." Thus, it would be difficult to say at exactly what point the standard-setting activity began. Certainly even the earliest development of the instrumentation was conducted in light of the eventual intent to create a summative evaluation which would distinguish among the typical, fully-competent classroom teacher (Level I), the outstanding classroom teacher (Level II), the truly distinguished classroom teacher (Level III).

A number of assumptions which related to the eventual standards had to be made early in the development process. First, teaching was viewed as belonging to the teacher rather than to the teacher-in-context. This was an essential consideration since if the quality of teaching did not generalize from one context to another (for example from one class to another or from one subject area to another), each teacher would have to be evaluated separately for each context in which the Career Ladder status would apply. The teacher-in-context could be expected to qualify for some settings but not for others. It was determined that the Career Ladder status should be assessed across the various contexts in which the candidate was currently performing (classes, subjects), and would continue to apply to

any grade level or subject area the teacher was entitled to teach for the duration of the certificate.

Second, the quality of teaching was viewed as a continuum. A Career Ladder III teacher was understood to be doing the same kinds of things as the Career Ladder I teacher, but to be doing them better. If this were not the case, a separate Career Ladder evaluation system would have been designed for each level of the Career Ladder.

These assumptions formed the basis of the evaluation system on which the standards would eventually be set. They became the "rules of the game." There were other rules as well. The intent of the Career Ladder system was to identify and reward meritorious teachers. It was clearly understood that this activity would require stringent standards. It was equally well understood that the standards were not to be based on a quota system. Any teacher who met the required standards was to be granted the appropriate status, regardless of how many teachers qualified. Data from the field test could be used as a guideline to establish the reliability and difficulty of the items and instruments, but in order to avoid quotas, standards had to be set in advance of any actual evaluations.

The evaluation system was designed in a criterion-referenced mode. A set of objectives was identified from the literature on effective teaching and then modified based on the professional judgment of Tennessee teachers. Once the content to be covered had been identified, it was organized into a set of Domains of Competence, subdivided into Indicators of Competence, and then further defined by measurement statements. These are presented in the Appendix. It was determined that because of the well-known flaws of various methods of evaluating teachers, it would be wise to measure each of the Domains and Indicators using several different sources of information. Instrumentation was, therefore, developed to assess each of the Domains and Indicators using at least three data sources. This led to a decision to base standards on the competencies rather than the data sources. It was considered more appropriate to have teachers receive scores in "teaching strategies" or "classroom management" than to have the scores associated with a portfolio or student questionnaires.

Because of the objectives-based approach, Tennessee was unable to locate any appropriate existing instruments. All of the instrumentation was, therefore, developed specifically for use with the Career Ladder evaluations. The instruments were pilot tested individually early in 1983-84 and then jointly field tested in the spring of 1984 on a randomly selected sample of Tennessee teachers. Fifty-two experienced and highly recommended teachers served as evaluators after two and a half weeks of

intensive training, which included reliability checks on every instrument. Over a period of six weeks, these field-test evaluators conducted evaluations of 276 teachers statewide. The data from this field test were an important source of information in the standard-setting process.

The data available to the standard-setting group was extensive. Score distributions for every item on every form (every scan on every observation, every student in every class) were available based on the teachers evaluated in the field test. In addition, evaluator judgment of every field-test candidate and every data source, together with extensive debriefing notes, were available, as well as information provided by the candidates themselves regarding the instruments and the evaluation process.

The instrumentation, process, and standards for the Career Ladder program were to be established by the Interim Certification Commission and approved by the State Board of Education. These bodies constituted the ultimate standard-setting panel. However, with the large amount of exceedingly complex information to be integrated into the standard-setting decision, it was deemed appropriate to select a separate panel of individuals who would have the time and expertise to fully understand the decisions to be made and to make recommendations on the scoring process and standards to be utilized in the Career Ladder evaluation system. The standard-setting Panel selected included field-test evaluators (individuals who were not only outstanding classroom teachers in their own right, but who had been trained in the evaluation system and had conducted the field-test evaluations) and members of the Interim Certification Commission, which had designed and approved the evaluation system. This Panel eventually met for almost a full day on each of four occasions. Department of Education staff members provided data analysis and other support services, but did not participate in the decision-making process.

The first decision to be made by the Panel was whether or not to retain each of the field-test instruments. Evaluator and candidate reactions to each instrument were reviewed, along with appropriate analyses of the field test data. In most cases, descriptive data were presented along with reliability analysis, correlational studies, factor analysis, and other relevant information (for example, the amount and distribution of missing data). The Panel eventually decided to retain all of the field-test instruments. However, it restructured the principal interview to become a questionnaire and required development and pilot testing of a K-2 student questionnaire. The latter was later approved for use in the system.

Having decided to retain the instruments, the Panel also made decisions on whether to drop individual items on them. The

decision to drop an item was made based on a combination of information, including reliability, redundancy, relationship to the appropriate indicator, degree of missing data, and clarity of purpose. A few items received minor rewording to improve clarity. No items were added to the actual evaluation, although several which appeared to be valuable but in need of substantial reworking were redone and included for field-test purposes only (not to be scored).

Finally, the panel decided on the first step in data reduction: aggregation. Items were aggregated over multiple respondents (students, peers), and also over multiple occasions (scans within observation visits, visits over candidates). Since this left more than 350 separate scores for each candidate, the Panel selected additional aggregations corresponding to the indicators within the Domains of Competence. These aggregations were used for all instruments except the observation where multiple scores within the same indicator were kept in some cases in order to improve conceptual clarity. In all, the Panel ended up with 85 scores (later reduced to 63 when the interview was dropped).

Staff then reanalyzed the data by combining individual items in the aggregation patterns selected. These scores were reviewed by the Panel at the next meeting. At this meeting, the Panel assigned weights to each of the scores within a domain in order to be able to generate domain scores which were as valid as possible. Comments by evaluators and candidates were combined with data on the individual items and professional judgment to determine the relative importance of the competency and the adequacy of its measurement. These considerations were debated extensively by the Panel members before they determined the table of weights to apply to the scores. The Panel also decided that separate scores should be calculated for each of the five domains for each candidate.

In order to be able to combine the scores using the weights, however, it was necessary to convert them to the same scale. The raw scores were on a variety of scales ranging from 1 through 5 on the questionnaires and 0 through 1 on portions of the observation, to actual numbers of students off-task or academic questions asked. As a first step, those scores which were dependent on variables (for example, the number of students in the classroom or the number of observation scans actually conducted) had already been standardized as percentages.

It was noted that most of the scores had skewed distributions (positive skew for most observation items, negative

for most items based on questionnaire data). After extensive discussion on the merits of various scales, a conversion to a nine-point scale, modeled loosely on a stanine, was adopted. In the process, scores for which a low score was preferable (e.g., classroom disruptions) were inverted so that a high score would always be a nine, and a low score a one.

Originally, the Panel members had hoped to set the cutoff scores on the nine-point scale at 8 for Level III and 7 for Level II. Inspection of the field-test data revealed that these cutoff scores were likely to be too high and would possibly result in few if any teachers qualifying. The data also indicated that performance had been substantially higher in some domains than in others and that individual teachers' performances had been very uneven, with none of the field-test candidates scoring high in all five domains.

Based on this information, the Panel set cutoff scores differentially in the five domains, with the cutoff highest in the planning domain where the highest field-test scores had been achieved. They also decided that, while for Level II the candidates would have to achieve the Level II (lower) cutoff in each of the five domains, for Level III the candidate would need to achieve the Level III (higher) cutoff score in only four of the five domains, provided that the fifth score met or exceed the Level II cutoff.

These recommendations, provided by the Standards Panel after many hours of deliberation, were presented to and approved by the Evaluation Committee of the Interim Certification Commission, the Interim Certification Commission and the State Board of Education. They were then published in the teacher orientation manual, which was distributed prior to any evaluations taking place.

Although many Tennessee educators still do not believe it, there was no quota system. Even those who scored the data had no information on how many teachers would qualify until after the last evaluation was complete and the results were scored for reporting purposes. This was possibly the highest risk that could have been taken. If few or no teachers had qualified, the credibility of the system would have been ruined. If all or nearly all teachers had qualified, the credibility of the system would also have been severely tested.

As it turned out, of the original group of 3,132 teachers completing the evaluation process, 482 (15%) qualified for Level II and 671 (21%) qualified for Level III. This was a surprise to everyone since it had been expected that more teachers would qualify for Level II than for Level III. (Many additional

teachers have qualified since that time and over 53% of this original group is now at Career Level II or III.) Additional surprises were the large number of teachers who exceeded the cutoff scores in all five domains, and the additional large number who qualified in every domain except Planning. The candidates' scores were, as expected, substantially higher in every Domain than they had been on the field test, but the higher cutoff scores in the Planning domain had nevertheless exceeded the abilities of more candidates than had the cutoff scores in the other areas.

It was also decided to set cutoff scores for each of the five scored domains since the committee felt that a high score in one domain of competence (for example, teaching strategies) could not compensate for a low score in another (for example, evaluation). This decision was particularly difficult since the committee felt strongly that certain domains, particularly teaching strategies and classroom management, were more important than the others. However, because the task was to identify outstanding rather than simply competent teachers, the committee felt it appropriate to require a high level of proficiency in all areas.

Again, the field-test data were reanalyzed, this time by scaling it to the nine-point scale as well as applying the appropriate weights. Since no other criterion score was available, evaluator direct judgment was used to divide the field-test candidates into three groups: those below Level II, those at Level II, and those at Level III. Score values for each group on each domain, along with overall score distributions for all candidates, were provided to the committee.

At this meeting, the discussion centered on what changes in scores could be expected from the field test to the actual evaluations. It was expected that a substantial increase in scores would occur. While the field-test candidates were a random sample of all teachers, the actual candidates would have a minimum of 8 years of teaching experience and be self-selected. While the field-test candidates had only a few weeks in which to prepare portfolios, the actual candidates would have many months. While the field-test candidates took the written test without any preparation or information on what the test would cover, the actual candidates would have study guides and months to prepare. While the field-test candidates were paid $100 for simply participating, the actual candidates had added incentives to perform as well as they could.

But knowing that performance would be better did not indicate how much better. There were no clear dividing lines between the scores of field-test candidates designated by evaluators as performing at levels I, II or III. Although the

means were in the correct order, there was overlap in the score distributions. Nor was it clear whether the lack of agreement was due to deficiencies on the part of the instruments, on the part of the evaluators, or simply because of the conditions of the field test itself.

It was clear in the legislation that the evaluation system was intended to be stringent, but due to the "no quotas" provision, no guidelines were available to define stringency. Since all of the instruments were designed for the system and had not been used elsewhere, no comparative data were available. For the good of the system, the standards needed to be high enough so that not everyone could qualify. They also needed to be low enough that some teachers could make the grade, and all could aspire to it.

Since the development and standard-setting for the teacher evaluation system of the Career Ladder program, a great deal has happened. After its first year of use, improvements were made in the teacher system for the 1985-86 year. The interview and portfolio were replaced by a synthesis of these data sources called a "Dialogue." The stanine-like scale was replaced by a standard score (z-score) scale for calculations and a 200-800 reporting scale with standards fixed at 600 for Level II and 700 for Level III. The standards are being maintained from the previous year's system by applying a variation on common item methodology to the individual Domains across data sources. A weighted total score is being used for qualification, with minimum scores of 450 required for each Domain. These minimums are based on a conversion for the average score attained by a Career Ladder I teacher in the previous year.

Evaluation systems have been developed and standards set for principals and assistant principals, supervisors of instruction, special education teachers who do not teach self-contained classes, vocational education taechers, teachers of the gifted, Chapter 1 and resource teachers, librarians and media specialists, and counselors. In each case, the standard-setting procedures were similar. The outcomes varied.

The special education and vocational education teacher systems are being linked to the standards of the teacher system using the same common item methodology employed to move the standards from 1984-85 to 1985-86. The same methodology is also being used to bring forward the principal/assistant principal and instructional supervisor standards established last year. Librarians/media specialists and counselors have established independent standards using the new metric. As previously, these standards have been established using field-test information and do not include any quotas.

Working with these systems, a great deal has been learned about the processes and outcomes of standard-setting under high-stakes conditions. The following are guidelines which are recommended to others coping with similar tasks:

1. Incorporate the standard-setting goal in the design, development, and field testing of the instrumentation. Good standards cannot stand alone. They must be based on good instruments.

2. Select knowledgeable, respected, and professional individuals for the standard-setting Panel. They need to be able to focus on the goal of the evaluation system amid a maze of complex considerations, to make difficult decisions, and to disagree with one another without being disagreeable. Once they have set the standards, they may also be called on to defend them.

3. Allow plenty of time. We have yet to complete this process in less than four meetings, with time between for additional data analysis and other information collection. If the Panel is not already intimately familiar with the evaluation system, it is essential that they become so before they attempt to set standards on it.

4. Use meaningful metrics wherever possible. We now present our standard-setting panels with the same data in the form of raw scores, percentiles, and percent correct. This is particularly valuable in preventing, for example, setting standards too near the floor or ceiling of the measurement range (percent correct), or so high or low that the standard becomes meaningless (percentile), or in a way which is inconsistent with the content of the score (raw score).

5. Keep clearly in mind the objective(s) of the standard. If the standard is intended to motivate improved performance, for example, it must be attainable.

6. Develop a plan for clearly communicating the standards and the process used to arrive at them. Work at building trust in the process. This is especially critical if the standards,

the metric, or any other aspect of the process is difficult for the candidates to comprehend.

7. Try to maintain a balance between objectivity essential for good decisions, but it cannot make them for you. Objective methods work very well for objects. Teachers are not objects, they are subjects. No instrument can adequately capture excellence in teaching, and no completely objective method could adequately set standards for a system designed to assess it. Good standard setting in this circumstance is both an art and a science.

Appendix

Data Sources for

COMPETENCIES AND INDICATORS

	Classroom Observation	Portfolio	Applicant Interview	Principal Interview	Peer Questionnaire	Student Questionnaire	Written Test
I. Prepares for instruction effectively.							
A. Establishes appropriate instructional goals and related objectives consistent with the curriculum.		X	X	X			
B. Prepares instructional plans and materials incorporating principles of effective instruction.	X	X	X	X			
C. Creates, selects or modifies instructional plans and materials to accommodate learner instructional levels.	X						
II. Uses teaching strategies and procedures appropriate to the content, objectives, and learners.							
A. Provides a clear description of the learning task and its content.	X	X				X	
B. Monitors learner understanding and reteaches as necessary.	X	X	X			X	
C. Provides learners appropriate practice and review.	X	X	X			X	
D. Establishes and maintains learner involvement in the learning task.	X		X			X	
III. Uses evaluation to improve instruction.							
A. Uses information about learner performance to improve the instructional process.		X	X	X			
B. Reports learner status and progress to learners and their parents.	X	X	X	X		X	
IV. Manages classroom activities effectively.							
A. Establishes and maintains appropriate learner behavior.	X	X		X		X	
B. Establishes and maintains a classroom climate conducive to learning.	X	X	X	X		X	
C. Makes effective use of classroom resources (e.g. personnel, time, materials, facilities)	X		X	X		X	
V. Establishes and maintains a professional leadership role.							
A. Improves professional skills and knowledge.		X	X	X			
B. Takes a leadership role in improving education.		X	X	X	X		
C. (Screening only) Performs professional responsibilities efficiently.					X		
VI. (Screening only) Communicates effectively.							
A. Writes clearly and correctly.	X					X	X
B. Communicates oral information effectively.	X					X	
C. Reads professionally relevant literature/materials with comprehension.							

BEST COPY AVAILABLE